DOCUMENT RESUME

ED 334 243 TM 016 788

AUTHOR De Ayala, R. J.; And Others
TITLE An Investigation of the Robustness of a Partial Credit Model-Based Computerized Adaptive Test to Misfitting Items.
PUB DATE Apr 91
NOTE 19p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Equations (Mathematics); *Error of Measurement; Factor Analysis; Goodness of Fit; Mathematical Models; Maximum Likelihood Statistics; *Robustness (Statistics); *Test Items
IDENTIFIERS Ability Estimates; *Partial Credit Model

ABSTRACT

The robustness of a partial credit (PC) model-based computerized adaptive test's (CAT's) ability estimation to items that did not fit the PC model was investigated. A CAT program was written based on the PC model. The program used maximum likelihood estimation of ability. Item selection was on the basis of information. The simulation terminated when a maximum of 30 items was reached or when a predetermined standard error of estimate (SEE) was obtained. SEE termination criteria of 0.20, 0.25, and 0.30 were used. Responses to 150 5-alternative items generated according to a linear factor analytic model were simulated for 1,000 examinees. Results indicate that reasonably accurate ability estimation could be obtained despite the adaptive tests, which, on the average, contained up to 45% misfitting items. The inclusion of the misfitting items did not appear to increase the PC CAT test lengths. The benefits of polytomous model-based CATs were discussed. Three data tables, five figures, and a 30-item list of references are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED334243

An Investigation of the Robustness of a Partial Credit Model-Based Computerized Adaptive Test to Misfitting Items

R.J. De Ayala,
University of Maryland

Barbara G. Dodd and William R. Koch,
University of Texas at Austin

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RALPH DE AYALA

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Please direct correspondence to:
R.J. De Ayala
Measurement, Statistics, and Evaluation
Benjamin Building
University of Maryland
College Park, MD 20742

Paper presented at the Annual Meeting of the NCME

TM016788

BEST COPY AVAILABLE

ABSTRACT

Computerized adaptive testing (CAT) is a procedure for administering tests which are individually tailored for each examinee. Although the majority of CATs are based on dichotomous item response theory (IRT) models, some researchers have explored the use of polytomous IRT models, such as the graded response model and partial credit (PC) model, in CAT. This study investigated the robustness of a PC model-based CAT's ability estimation to items which did not fit the PC model. Results showed that for the PC CAT, reasonably accurate ability estimation ($r_{\hat{\theta}\theta_T} \geq 0.921$) may be obtained despite adaptive tests which, on average, contained up to 45% misfitting items. Furthermore, the inclusion of misfitting items did not appear to increase the PC CAT test lengths. The benefits of polytomous model-based CATs were presented.

One important and very promising application of item response theory (IRT) is computerized adaptive testing (CAT). Unlike the conventional paper-and-pencil test in which an examinee is administered all test items, CAT is a procedure for administering tests which are individually tailored for each examinee. The advantage of IRT-based CAT over paper-and-pencil testing have been well documented (e.g., Weiss, 1982). Although not necessary (cf., De Ayala, Dodd, & Koch, 1990), a CAT system typically uses an IRT model in combination with test item characteristics to estimate the examinee's ability. Typically, either the three-parameter logistic or Rasch models (e.g., McBride & Martin, 1983; Kingsbury & Houser, 1988) have been used in CAT. Despite research which has demonstrated the existence of partial knowledge of the correct answer (e.g., Levine & Drasgow, 1983; Thissen, 1976), dichotomous models and dichotomous model-based CATs operate as if an examinee either knows the correct answer or randomly selects an incorrect alternative.

Some research has explored the benefits and operating characteristics of CATs based on polytomous IRT models (e.g., De Ayala, 1989; Dodd, Koch, & De Ayala, 1989; Koch & Dodd, 1989; Sympson, 1986). In general, these studies have shown that item pools smaller than those used with dichotomous model-based CATs have led to satisfactory estimation, that the use of the ability's standard error of estimation for terminating the adaptive test is preferred to the minimum item information termination criterion, and the use of a variable stepsize instead of a fixed stepsize tends to minimize nonconvergence of trait estimation. In addition, it should be noted that polytomous model-based CAT may be used not only with polytomously scored items, but with solely dichotomously scored items, or with a combination of the two (i.e., some items are scored polytomously while others are scored dichotomously).

Polytomous graded models have been used for the assessment of the clinical competence of physicians (Julian & Wright, 1988), the construction and analysis of writing tests (Ackerman, 1986; Pollitt and Hutchinson, 1987), educational diagnosis (Adams, 1988), and in CAT for the administration of Likert-type attitude questions and personality inventories (Koch & Dodd, 1985; Dodd, 1985; Koch, 1983). Given that, a number of aptitude test items have traditionally been scored in a graded fashion it is reasonable and desirable to expect that CAT implementations in these subjects to incorporate a graded scoring system. For instance, statistics, chemistry, and physics exams are typically graded by giving partial credit for some incorrect answers. Therefore, it would appear reasonable to expect that the use of partial credit scoring for some incorrect answers would enhance the acceptance of CAT in these area. Three polytomous graded models whose properties for CAT have been studied are Samejima's (1969) graded response model, the rating scale model (Andrich, 1978), and Masters' (1982) partial credit (PC) model (e.g., Dodd, Koch, & De Ayala, 1989; Koch & Dodd, 1989; Dodd, Koch, & De Ayala, in press).

To obtain the advantages of the PC model (and IRT models in general) there must be satisfactory model-data fit. To the extent that there is low model-data fit, some or all of the advantages of the model may be lost. Although the assessment of model-data fit may be approached via a number of different techniques (cf., Hambleton & Rogers, 1986; Ludlow, 1986; Kingston & Dorans, 1985; Wright & Masters, 1982; Yen, 1981), one common approach is to use fit statistics.

The Rasch perspective involves retaining only those items which are found to fit the model. Strictly speaking, items which do not fit the model are examined to determine the cause of misfit and may still be retained if it is felt that the misfit is due to a few large residuals. Calibration programs for the Rasch family of models traditionally output a number of fit statistics, as well as information from other model-data fit approaches.

Although Koch and Dodd (1989) and Dodd, Koch, and De Ayala (1989) have investigated various facets of adaptive testing with the PC model (i.e., item pool size, stepsizes, information functions), one factor which has not been addressed and which is crucial for any implementation is the robustness of the PC model-based CAT to violations of data fit. Because the creation of the item pool involves the interaction of the subjective interpretation of model-data fit as well as logistical and administrative factors, the item pool will consist of items which will vary in their degree of fit (or misfit). For instance, items may be included in an item pool for reasons of content validity (although the items may not fit well). Therefore, this study addressed how robust was the PC model-based CAT's ability estimation to the use of items which did not fit the models.

MODEL

The PC model is appropriate for items with ordered responses, such as aptitude and achievement test items whose alternatives are inherently ordered or have been ordered according to degree of correctness (e.g., through partial credit scoring). In addition, attitude questionnaires and ratings data may also be fitted by the model.

The PC model provides a direct expression of the probability of an examinee with ability θ responding in a particular category. In the PC model the examinee-item interaction is modeled as :

$$p_{x_i}(\theta) = \frac{e^{\sum_{j=0}^{x_i}(\theta - b_{x_i})}}{\sum_{k=0}^{m_i} e^{\sum_{j=0}^{k}(\theta - b_{x_i})}} \qquad (1),$$

where θ is the latent trait, b_{x_i} is the difficulty parameter of the step associated with the category score x_i; item i has m_i categories and $x_i = 1..m_i$. A category score reflects the number of successfully completed steps. A "step" is simply a stage required to complete an item. For instance, the problem $((6/3)+2)^2$ is considered to contain three steps because there are three separate stages which must be completed (in a specific order) to correctly answer the problem (i.e., step 1 : 6/3, step 2 : the addition of 2 to the quotient, and step 3 : the squaring of the quantity). For notational convenience $\Sigma(\theta - b_{x_i})$ where $j=0$ is defined as being equal to zero.

Because the PC model is an extension of the Rasch model it assumes that all items are equally good at discriminating among examinees. In addition, as a member of the Rasch family, the PC model's item and person parameters may be estimated on the basis of the existence of sufficient statistics. Specifically, an examinee's test score contains all the information for estimating his or her ability and the items' difficulties may be estimated from a simple count of the number of persons completing each "step" of an item. The PC model requires that the steps within an item be completed in sequence, although the steps need not be equally difficult nor be ordered in terms of difficulty. If an item consists of only two categories, then the PC model reduces to the Rasch model.

METHOD

Programs: A CAT program was written based on the PC model (PC CAT). The program used maximum likelihood estimation (MLE) of ability and item selection was on the basis of information. The adaptive testing simulation was terminated when either of two criteria

were met : a maximum of thirty items was reached or when a predetermined standard error of estimate (SEE) was obtained (SEE termination criteria of 0.20, 0.25, 0.30 were used). Previous research with polytomous model-based CATs has shown that SEE results in better CAT performance than does the minimum item information criterion (e.g., Dodd, Koch, & De Ayala, 1989). The initial ability estimate for an examinee was the population's mean and a variable stepsize was used for ability estimation when MLE was not possible.

Data : One thousand simulees were randomly selected from a N(0,1) distribution (the z-scores were considered to be the simulees' true ability, θ_T). The examinees' responses to 150 5-alternative items generated according to a linear factor analytic model (Wherry, Naylor, Wherry, & Fallis, 1965) in which :

$$z_{ij} = a_j z_i + \sqrt{1 - h_j^2} z_{e_{ij}} \qquad (2),$$

where z_i was examinee i's randomly selected z-score (i.e., θ_T), a_j was item j's factor loading, h_j^2 was item j's communality, $z_{e_{ij}}$ was a z-score random number that was generated specifically for the error component of item j and examinee i. Subsequent to the calculation of z_{ij}, z_{ij} was compared to pre-specified category boundaries to determine the category response for examinee i to item j. All factor loadings were uniformly high and ranged from 0.62 to 0.85. The category boundaries used may be found in Dodd (1985).

The use of a linear factor analytic approach for data generation allowed item discriminations to vary and the responses to be a nonogival function of ability (i.e., a violation of a fundamental IRT assumption).

Calibration: MSTEPS (Wright, Congdon, & Schultz, 1989) was used to obtain item parameter estimates and fit statistics for the PC model.

Fit Analysis: For the purpose of this study the weighted total fit statistic was chosen for identifying item misfit for the PC model; the weight is the information function and is used to reduce sensitivity to outliers (Smith, 1988).

The original 1000 x 150 data matrix was calibrated and fit statistics were obtained. After the elimination of items deemed to show "significant" misfit, the data set was recalibrated without the misfitting items. Fit was then reexamined and items found to fit were retained; their item parameter estimates were used for the item pools. Because model-data fit is a matter of degree, various critical values (CV) were used to determined whether an item was exhibiting significant misfit. For the PC model the CVs used were ± 2.0, ± 3.0, ± 4.0, ± 5.0 (roughly corresponding to α values of 0.046, 0.003, less than 0.0001, less than 0.0001, respectively) and the CVs = $\pm \infty$ (i.e., all items were considered to fit and included in the item pool).

Summary: A 1000 examinee by 150 item data matrix was generated and calibrated. Critical values (5 levels) were used for identifying misfitting items. Subsequent to the elimination of misfitting items, the data were recalibrated and reexamined for misfit. When no items were found to misfit, the item parameter estimates were used to create a CAT item pool; five item pools for the PC CAT were created (one corresponding to each CV level for each model). The design consisted of the crossing of the SEE factor (3 levels : 0.20, 0.25, 0.30) by the CV factor. For each of the 1000 examinees an adaptive test was simulated using each item pool for the PC CAT.

Analysis: The CAT simulations were analyzed by comparing each CAT's estimated ability ($\hat{\theta}$) with θ_T through correlational analysis (Pearson product-moment correlation coefficients: $r_{\hat{\theta}\theta_T}$), average absolute differences (AAD), standardized root mean squared differences (SRMSD) and standardized differences between means (SDM) (Doody-Bogan and Yen, 1983) where :

$$AAD = \frac{\sum_{i=1}^{N} |\hat{\theta}_j - \theta_{T_j}|}{N} \qquad (3)$$

$$SRMSD = \sqrt{\frac{\frac{1}{N}\sum_{j=1}^{N}(\hat{\theta}_j - \theta_T)^2}{\frac{s^2_{\hat{\theta}} + s^2_{\theta_T}}{2}}} \qquad (4)$$

$$SDM = \frac{\bar{\hat{\theta}} - \bar{\theta}_T}{\sqrt{\frac{s^2_{\hat{\theta}} + s^2_{\theta_T}}{2}}} \qquad (5),$$

where $\hat{\theta}_j$ was the ability estimate for examinee j, θ_{T_j} was the known true ability for examinee j, N was the number of examinees, $\bar{\theta}_T$ was the mean θ_T, $\bar{\hat{\theta}}$ was the mean of $\hat{\theta}$, $s^2_{\hat{\theta}}$ was the variance of $\hat{\theta}$, $s^2_{\theta_T}$ was the variance of θ_T. The differences between $\hat{\theta}$ and θ_T as a function of θ_T were graphically examined (a.k.a., difference plots). Further, descriptive statistics were calculated on the number of items administered, the item pools, the proportion of misfitting items administered relative to the use of the most conservative CV was obtained (i.e., CV = ±2.0), and the item pools' estimated information functions was inspected.

RESULTS

Calibration and Fit Analysis

For the PC calibration 33, 51, 63, and 78 items were found to fit the PC model using the CVs of ±2.0, ±3.0, ±4.0, ±5.0, respectively. The nomenclature for the corresponding item pools is : model + the number of items in the pool (e.g., PC 33 is the pool for the PC model containing 33 items and based on CV = ±2.0).

The PC 33-, 51-, 63-, 78-, and 150-item pools had step difficulty estimates which ranged from -2.50 to 3.03, -2.38 to 3.14, -2.35 to 3.13, -2.44 to 2.97, -3.0 to 3.31, respectively. Figure 1 shows the total item pool information for the PC 33-, 51-, 63-, and 78-item pools.

Insert Figure 1 about here

CAT Simulations

For the PC CAT simulations the correlation coefficients between $\hat{\theta}$ and θ_T increased as the SEE termination criterion decreased (see Table 1). All correlation coefficients were equal to or greater than 0.87 and the corresponding scatterplots showed strong linear associations. As can be seen even with the 33 item pool there was a strong linear association between $\hat{\theta}$ and θ_T. Becoming less conservative with respect to the magnitude of the CV (up to about ±4.0) produced $r_{\hat{\theta}\theta_T}$s of more or less comparable magnitudes to those obtain with CVs of ±2.0 and an increase in the number of examinees whose ability estimates were considered reasonable.

Insert Table 1 about here

Difference plots (i.e., $\hat{\theta} - \theta_T$ as a function of θ_T) for selected PC CATs are presented in Figure 2; these plots are typical of all the PC CAT plots. As can be seen the PC CATs did not tend to either underestimate or overestimate θ_T in a systematic way. In general, as SEE termination criterion decreased the points tended to become less variable about the baseline of 0.

Insert Figure 2 about here

AAD and SRMSD provide an assessment of the accuracy of estimation across examinees, while SDM assesses the overall bias between the $\hat{\theta}$s and θ_Ts. The SRMSD and SDM for the PC CATs are presented in Table 2. As can be seen, compared to the use of the ±2.0 CV, overall accuracy increased when the CVs of ±3.0 and ±4.0 were used. Regardless of the item pool used, the minimal bias exhibited by the PC CAT may not be considered

meaningful by some. Although SRMSD and SDM are aggregate indices and therefore, compensation may occur, the difference plots and the AAD indices showed that this was not the case. The AAD indices reflected the SRMSD/SDM pattern, that is, CVs of ±3.0 and ±4.0 resulted in the smallest AAD.

Insert Table 2 about here

Table 3 contains descriptive statistics on the PC adaptive tests. As would be expected, decreasing the SEE termination criterion produced an increase in average and median test lengths. Similarly, decreasing the SEE termination criterion resulted in an increase in the proportion of misfitting items administered. Comparing Tables 1 and 3, one sees that $r_{\hat{\theta}\theta_T} = 0.963$ and $r_{\hat{\theta}\theta_T} = 0.959$ were obtained (based on 98.8% and 99.0% of the examinees, respectively) despite the administration of tests containing, on average, 35.4% (CV = ±3.0) and 45.5% (CV = ±4.0) misfitting items. Inspection of plots of the proportion of misfitting items administered versus θ_T showed no systematic relationship.

Insert Table 3 about here

DISCUSSION

Using a CV = ±2.0 only 22% of the original items were found to fit the PC model. As stated above, each of the 117 items which were found to have significant fit statistics would have had to been analyzed separately to determine the cause of the misfit. For instance, the 1000 examinees could be ordered by their ability and their responses examined to see if individuals with abilities above and below the item's location were behaving according to expectations. If the majority of the examinees were behaving according to how the model would predict they should and the fit statistic's significance could be attributed to discrepancies in the expectations of a few examinees, then the item would be retained and the analysis would proceed to the next misfitting item. Of course, with large numbers of examinees and a large number of misfitting items this procedure would be arduous at best. However, the results showed that strong linear associations could be obtained despite the inclusion items which did not fit the PC model at CV = ±2.0. In fact, when the entire item pool was used and with an SEE termination criterion of 0.20, then a fidelity coefficient of 0.945 with comparatively low AAD/SRMSD and SDM values was obtained. The tradeoff for being able to include a large number of misfitting items was a substantial increase in the number of individuals whose $\hat{\theta}$s were not considered reasonable (i.e., $\hat{\theta} \leq -4.0$ or $\hat{\theta} \geq 4.0$).

Given the $r_{\hat{\theta}\theta_T}$s, the difference plots, SRMSD, SDM, and AAD results for the PC CAT, it appears that item pools smaller than are suggested for dichotomous model-based CATs can

be used with PC model-based CATs; this result replicates Dodd, Koch, & De Ayala (1989) and Koch and Dodd's (1989) findings. It appears that reasonably accurate ability estimation may be obtained despite adaptive tests which, on average, contained up to 45% misfitting items (i.e., the use of CV = $\pm$ 4.0 or less). Furthermore, the inclusion of misfitting items did not appear to increase the PC CAT test lengths.

References

Ackerman, T. (1986, April). *Use of the graded response IRT model to assess the reliability of direct and indirect measures of writing assessment.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Adams, R.J. (1988). Applying the partial credit model to educational diagnosis. *Applied Measurement in Education, 1*, 347-362.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

De Ayala, R.J. (1989). Computerized adaptive testing: A comparison of the nominal response model and the three-parameter model. *Educational and Psychological Measurement, 49*, 789-805.

De Ayala, R.J., Dodd, B.G. & Koch, W. R. (1990). A computerized simulation of a flexilevel test and its comparison with a Bayesian computerized adaptive test. *Journal of Educational Measurement, 27*, 227-239.

Dodd, B.G. (1985). *Attitude scaling : A comparison of the graded response and partial credit latent trait models* (Doctoral dissertation, University of Texas at Austin, 1984). *Dissertation Abstracts International, 45*, 2074A.

Dodd, B.G., Koch, W.R., & De Ayala, R.J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13*, 129-144.

Dodd, B.G., Koch, W.R., & De Ayala, R.J. (in press). Computerized adaptive attitude measurement: A comparison of the graded response and rating scale models. *Applied Psychological Measurement.*

Doody-Bogan, E., & Yen, W.M. (1983, April). *Detecting multidimensionality and examining its effects on vertical equating with the three-parameter logistic model.* Paper presented at the annual meeting of American Educational Research Association, Montreal.

Hambleton, R.K., & Rogers, H.J. (1986, April). *Promising direction for assessing item response model fit to test data.* Paper presented at the meeting of the American Educational Research Association, San Francisco.

Julian, E.R., & Wright, B.D. (1988). Using computerized patient simulations to measure the clinical competence of physicians. *Applied Measurement in Education, 1*, 299-318.

Kingsbury, G.G., & Houser, R.L. (1988, April). *A comparison of achievement level estimates from computerized adaptive testing and paper-and-pencil testing.* Paper presented at the annual meeting of American Educational Research Association, New Orleans.

Kingston, N.M., & Dorans, N.J. (1985). The analysis of item-ability regressions : An exploratory IRT model fit tool. *Applied Psychological Measurement, 9*, 281-288.

Koch, W.R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement, 7*, 15-32.

Koch, W.R., & Dodd, B.G. (1985, April). *Computerized adaptive attitude measurement.* Paper presented at the meeting of the American Educational Research Association, Chicago.

Koch, W.R., & Dodd, B.G. (1989). An investigation of procedures for computerized adaptive testing using the partial credit scoring. *Applied Measurement in Education, 2*, 335-357.

Levine, M., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43*, 675-685.

Ludlow, L.H. (1986). Graphical analysis of item response theory residuals. *Applied Psychological Measurement, 10*, 217-229.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

McBride, J.R., & Martin, J.T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D.J. Weiss (ed.), *New Horizons in Testing* (pp 223-237). New York : Academic.

Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing, 4*, 72-92.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Smith, R.M. (1988, April). *A comparison of the power of Rasch total and between item fit statistics to detect measurement disturbances.* Paper presented at the meeting of the American Educational Research Association, San Francisco.

Sympson, J.B. (1986, August). *Extracting information from wrong answers in computerized adaptive testing.* Paper presented at the American Psychological Association, Washington, D.C.

Thissen, D.J. (1976). Information in wrong responses to Raven's Progressive Matrices. *Journal of Educational Measurement, 13*, 201-214.

Weiss, D.V. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

Wherry, R.J., Sr., Naylor, J.C., Wherry, R.J., Jr., & Fallis, R.F. (1965). Generating multiple samples of multivariate data with arbitrary population parameters. *Psychometrika, 30*, 303-314.

Wright, B.D., Congdon, R., & Schultz, M. (1989). *A user's guide to MSTEPS* (Version 2.4). Chicago : MESA Psychometric Laboratory.

Wright, B.D., & Masters, G.N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.

Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Table 1: Pearson product-moment correlation coefficients between $\hat{\theta}$ and θ_T for PC CAT.

	Fit Statistics				
SEE	±2.0	±3.0	±4.0	±5.0	All items
0.30	0.919	0.923	0.921	0.902	0.870
0.25	0.944	0.943	0.943	0.934	0.907
0.20	0.963	0.963	0.959	0.952	0.945
Pool Size	33	51	63	78	150
N[a]	958	988	990	857	737

[a]refers to the number of cases whose ability estimates fell within the range ±4.0

Table 2: SRMSD, SDM, and AAD for PC CAT

Fit Statistics	SEE	SRMSD	SDM	AAD
±2.0	0.30	0.471	-0.199	0.345
	0.25	0.419	-0.218	0.308
	0.20	0.363	-0.212	0.269
±3.0	0.30	0.405	-0.065	0.315
	0.25	0.356	-0.071	0.273
	0.20	0.295	-0.076	0.222
±4.0	0.30	0.406	-0.039	0.316
	0.25	0.353	-0.035	0.262
	0.20	0.304	-0.045	0.225
±5.0	0.30	0.501	-0.136	0.343
	0.25	0.437	-0.156	0.292
	0.20	0.390	-0.166	0.259
All items	0.30	0.628	-0.114	0.351
	0.25	0.528	-0.108	0.304
	0.20	0.396	-0.076	0.232

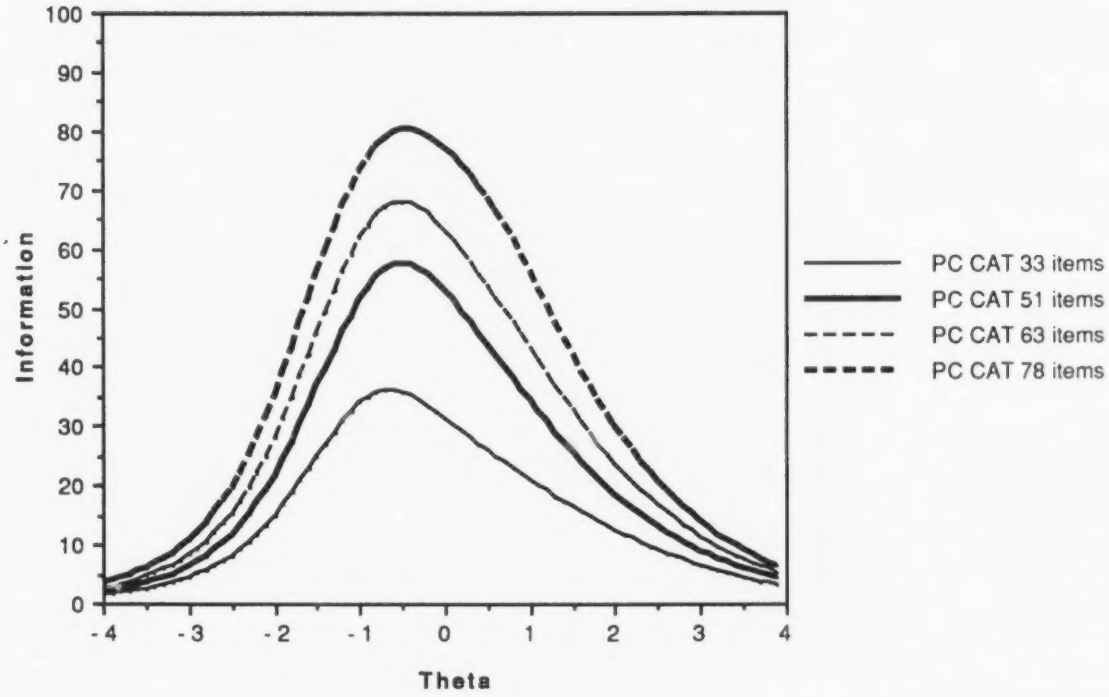
Table 3: Descriptive Statistics for PC CAT

Fit Statistics	SEE	Mean NIA[a]	Median NIA[a]	SD NIA[a]	Range	Proportion[b]
±2.0	0.30	8.56	7	3.33	6-30	-
	0.25	13.01	11	5.36	9-30	-
	0.20	21.63	20	5.96	14-30	-
±3.0	0.30	8.11	7	2.85	6-30	0.213
	0.25	11.79	10	4.22	9-30	0.288
	0.20	18.70	17	5.27	14-30	0.354
±4.0	0.30	7.89	7	2.65	6-30	0.375
	0.25	11.28	10	3.53	9-30	0.426
	0.20	17.88	16	4.77	13-30	0.455
±5.0	0.30	7.69	7	2.19	6-30	0.460
	0.25	10.98	10	2.94	9-30	0.504
	0.20	17.24	16	3.87	14-30	0.527
All items	0.30	7.98	7	2.53	6-26	0.637
	0.25	11.20	10	3.24	9-30	0.655
	0.20	17.31	16	3.83	13-30	0.671

[a]Number of items administered

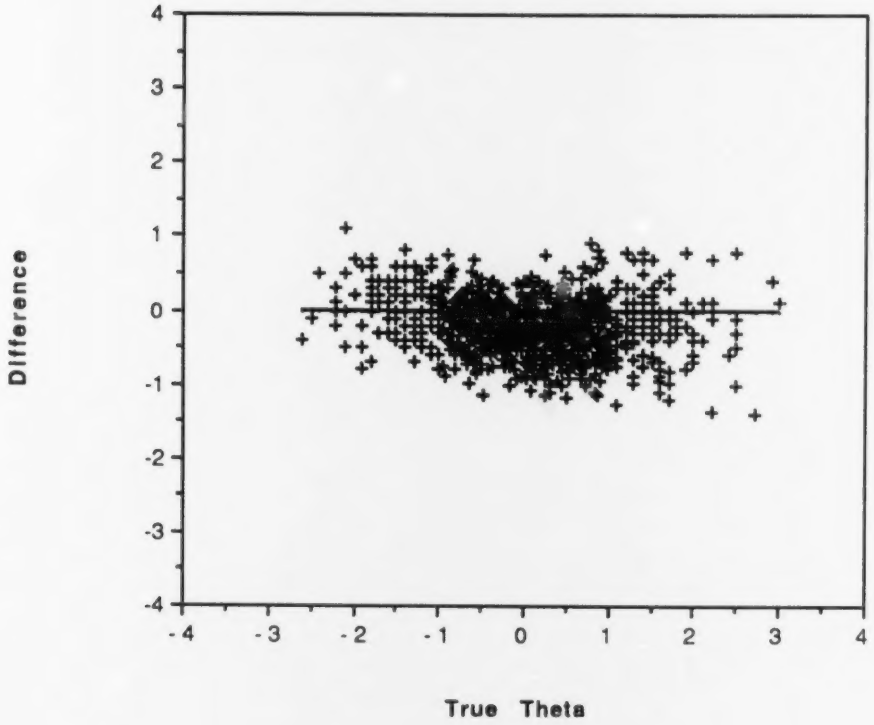
[b]Proportion of misfitting items administered relative to the use of CV = ±2.0

Figure 1. Information function estimates: PC model 33-, 51-, 63-, and 78-item pools



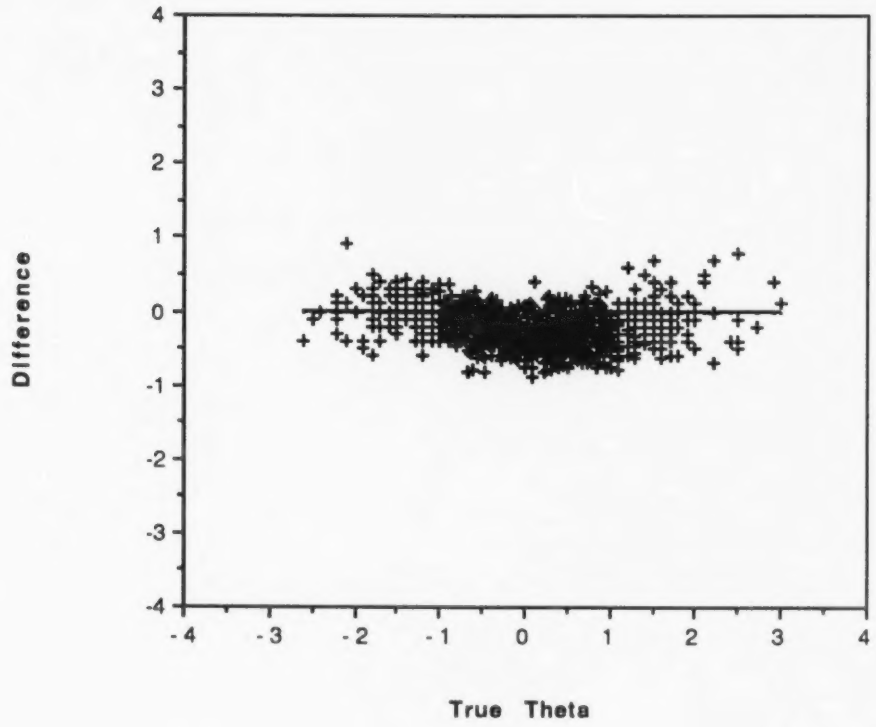
BEST COPY AVAILABLE

Figure 2a. Difference plots ($\hat{\theta}$ - θ_T) for the PC CAT: 33-item pool, termination SEE = 0.30



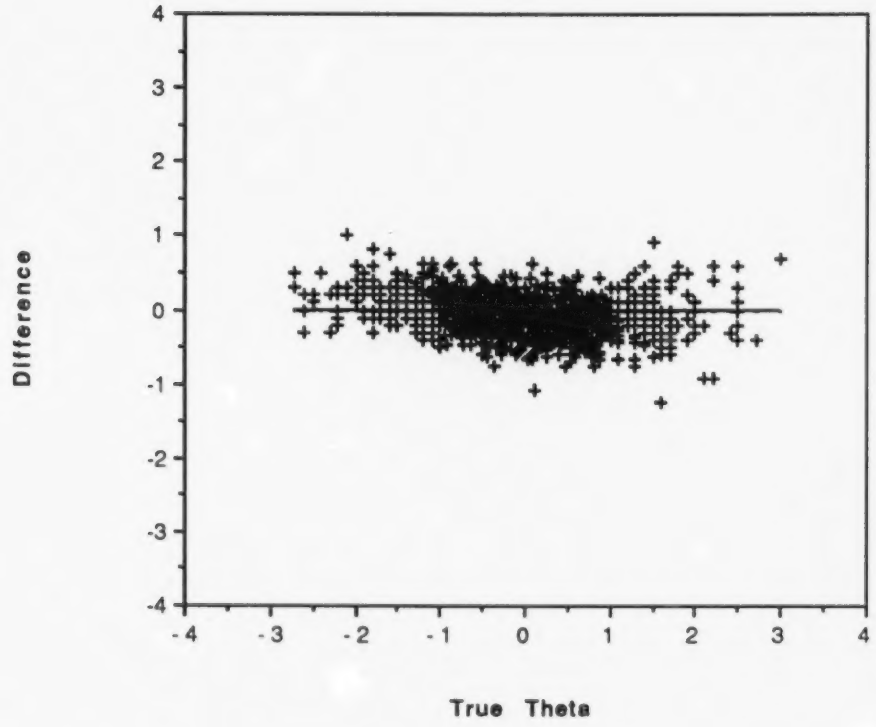
BEST COPY AVAILABLE

Figure 2b. Difference plots ($\hat{\theta} - \theta_T$) for the PC CAT: 33-item pool, termination SEE = 0.20



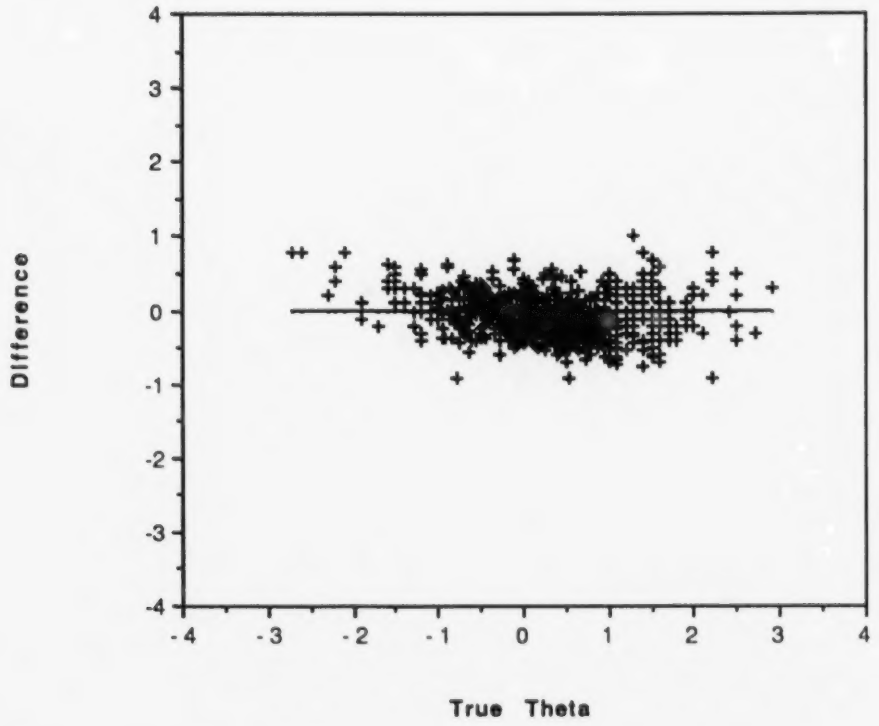
BEST COPY AVAILABLE

Figure 2c. Difference plots ($\hat{\theta}$ - θ_T) for the PC CAT: 63-item pool, termination SEE = 0.20



BEST COPY AVAILABLE

Figure 2d. Difference plots ($\hat{\theta}$ - θ_T) for the PC CAT: 150-item pool, termination SEE = 0.20



BEST COPY AVAILABLE